Generation and evaluation of synthetic patient data

Year: 2020 | Citations: 319 | Authors: André R. Gonçalves, P. Ray, Braden C. Soper, Jennifer L. Stevens, Linda Coyle

Abstract

Background Machine learning (ML) has made a significant impact in medicine and cancer research; however, its impact in these areas has been undeniably slower and more limited than in other application domains. A major reason for this has been the lack of availability of patient data to the broader ML research community, in large part due to patient privacy protection concerns. High-quality, realistic, synthetic datasets can be leveraged to accelerate methodological developments in medicine. By and large, medical data is high dimensional and often categorical. These characteristics pose multiple modeling challenges. Methods In this paper, we evaluate three classes of synthetic data generation approaches; probabilistic models, classification-based imputation models, and generative adversarial neural networks. Metrics for evaluating the quality of the generated synthetic datasets are presented and discussed. Results While the results and discussions are broadly applicable to medical data, for demonstration purposes we generate synthetic datasets for cancer based on the publicly available cancer registry data from the Surveillance Epidemiology and End Results (SEER) program. Specifically, our cohort consists of breast, respiratory, and non-solid cancer cases diagnosed between 2010 and 2015, which includes over 360,000 individual cases. Conclusions We discuss the trade-offs of the different methods and metrics, providing guidance on considerations for the generation and usage of medical synthetic data.